Classifying queries for people
from a people search engine
A people search engine
Popularity of people search

- One can search for:
  - Film stars
  - Politicians
  - School teachers
  - Friends
  - Yourself
  - ...

- An estimated 11-17% of web search queries contain a person name (2004 data)
Taxonomy

- Non-famous
- High-profile
  - Event-based
  - Regular
Non-famous
Event-based: “Ell Nikki”
Regular: “Geert Wilders”
Motivation

- Show different results
- Apply a different ranking algorithm
- Show results in a different way
Data

- 4 months from a Dutch people search engine:
  - 13 million searches
  - 4 million clickouts
- Annotated set: 216 total
  - 132 non-famous
  - 60 event-based
  - 24 regular
Features: Attention

Event-based

Regular
Attention = Volume

- Search volume, Clickout volume, News volume
  - Total
  - Last week
  - Trend
  - Burstiness
Other features

- Clickout entropy
  - over urls
  - over top level domains
- Wikipedia presence (Dutch)
  - Title match
  - Overall frequency
Experiment setup

- Two experiments:
  - 2-way: High vs Low
  - 3-way: Regular vs Event vs Low
- Stratified 10-fold cross-validation
- Classifiers:
  - C4.5
  - Naive Bayes
  - SVM
## Results: High vs Low

<table>
<thead>
<tr>
<th>Query Type</th>
<th>C4.5</th>
<th>Naive Bayes</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Profile</td>
<td>0.85</td>
<td>0.82</td>
<td>0.89</td>
</tr>
<tr>
<td>Low profile</td>
<td>0.89</td>
<td>0.91</td>
<td>0.81</td>
</tr>
</tbody>
</table>
## Results: Regular vs Event vs Low

<table>
<thead>
<tr>
<th>Query Type</th>
<th>C4.5 Prec.</th>
<th>C4.5 Rec.</th>
<th>Naive Bayes Prec.</th>
<th>Naive Bayes Rec.</th>
<th>SVM Prec.</th>
<th>SVM Rec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event based</td>
<td>0.83</td>
<td>0.87</td>
<td>0.74</td>
<td>0.62</td>
<td>0.85</td>
<td>0.55</td>
</tr>
<tr>
<td>Regular</td>
<td>0.57</td>
<td>0.54</td>
<td>0.53</td>
<td>0.33</td>
<td>0.45</td>
<td>0.38</td>
</tr>
<tr>
<td>Low profile</td>
<td>0.92</td>
<td>0.90</td>
<td>0.81</td>
<td>0.92</td>
<td>0.80</td>
<td>0.96</td>
</tr>
</tbody>
</table>
Decision tree C4.5
Error analysis

- Points misclassified – 33 out of 216 (C4.5)
- Annotator agreement – 0.70 (Cohen’s kappa)
  - disagreement 40 out of 216 or 19%
- Identified reasons for error for the 15 ‘hard’ points:
  - overlooking of evidence by an annotator
  - ambiguity
  - evidence not strong enough for classifiers
  - small size of annotated set
Future work

- Try to address the ambiguity problem (hard) using:
  - the document collection
  - evidence from the search logs
  - evidence from the click logs
  - external evidence